

Data Mining

Cluster Analysis: Advanced Concepts and Algorithms

Lecture Notes for Chapter 9

Introduction to Data Mining
by
Tan, Steinbach, Kumar

Edited by J. Taylor for STATS202, Stanford University, Winter 2009

Fuzzy clustering

- Each point belongs to j -th cluster with weight w_{ij}
- Minimizes, with constraints on sum of each cluster's w 's=1

$$SSE(C) = \sum_j \sum_i w_{ij}^p \text{dist}(x_i, c_j)^2$$

- Advantages: uses all points for each cluster
- Disadvantages: same as K-means.

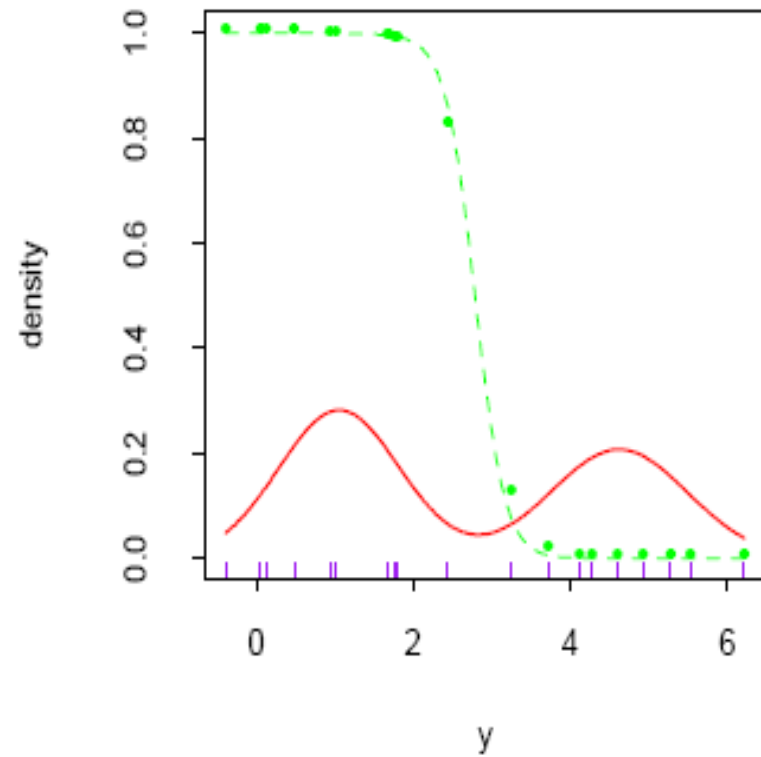
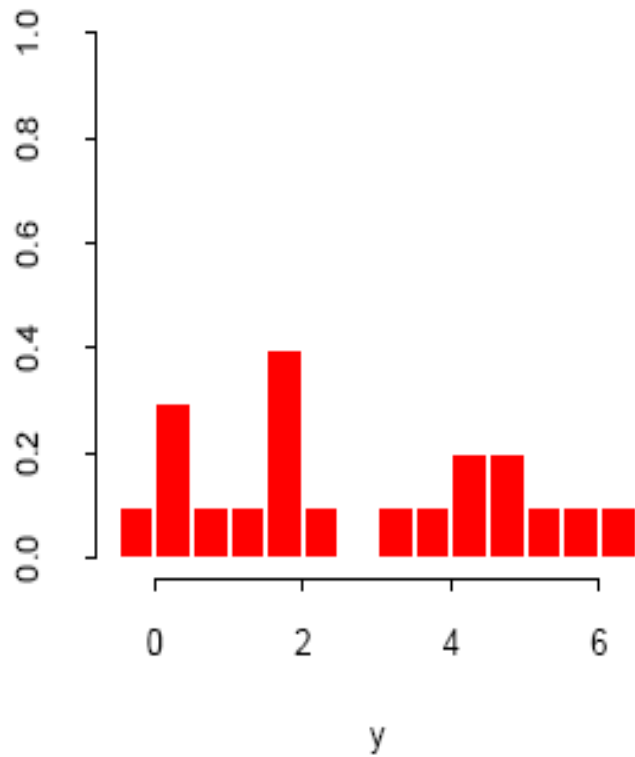
Mixture modelling

- Consider each observation as coming from a mixture of distributions

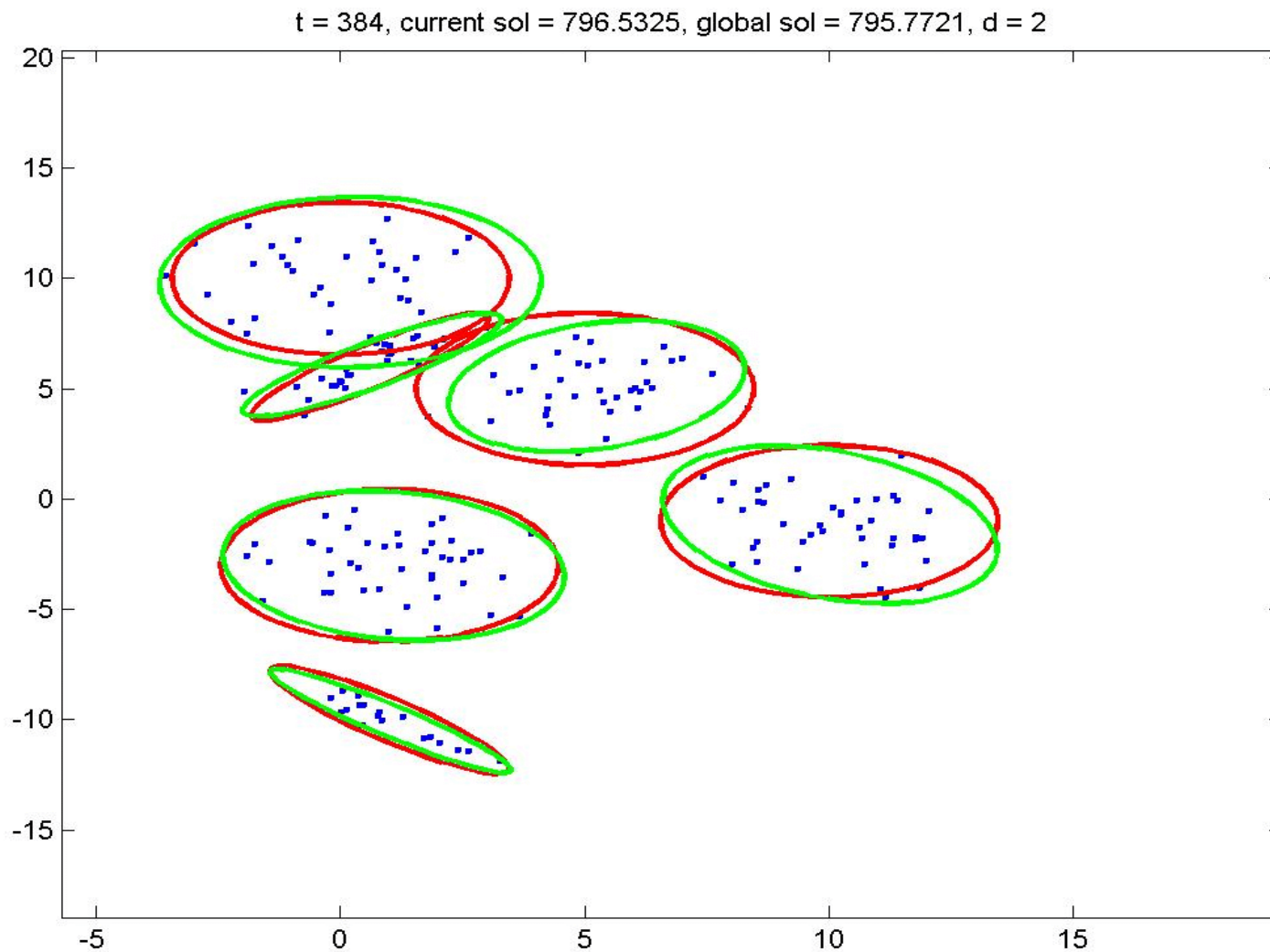
$$f(x) = \sum_j \pi_j f(x; \theta_j)$$

- Common example: components are Gaussian.
- Algorithm to fit model: EM algorithm.

Mixture modelling



Mixture modelling



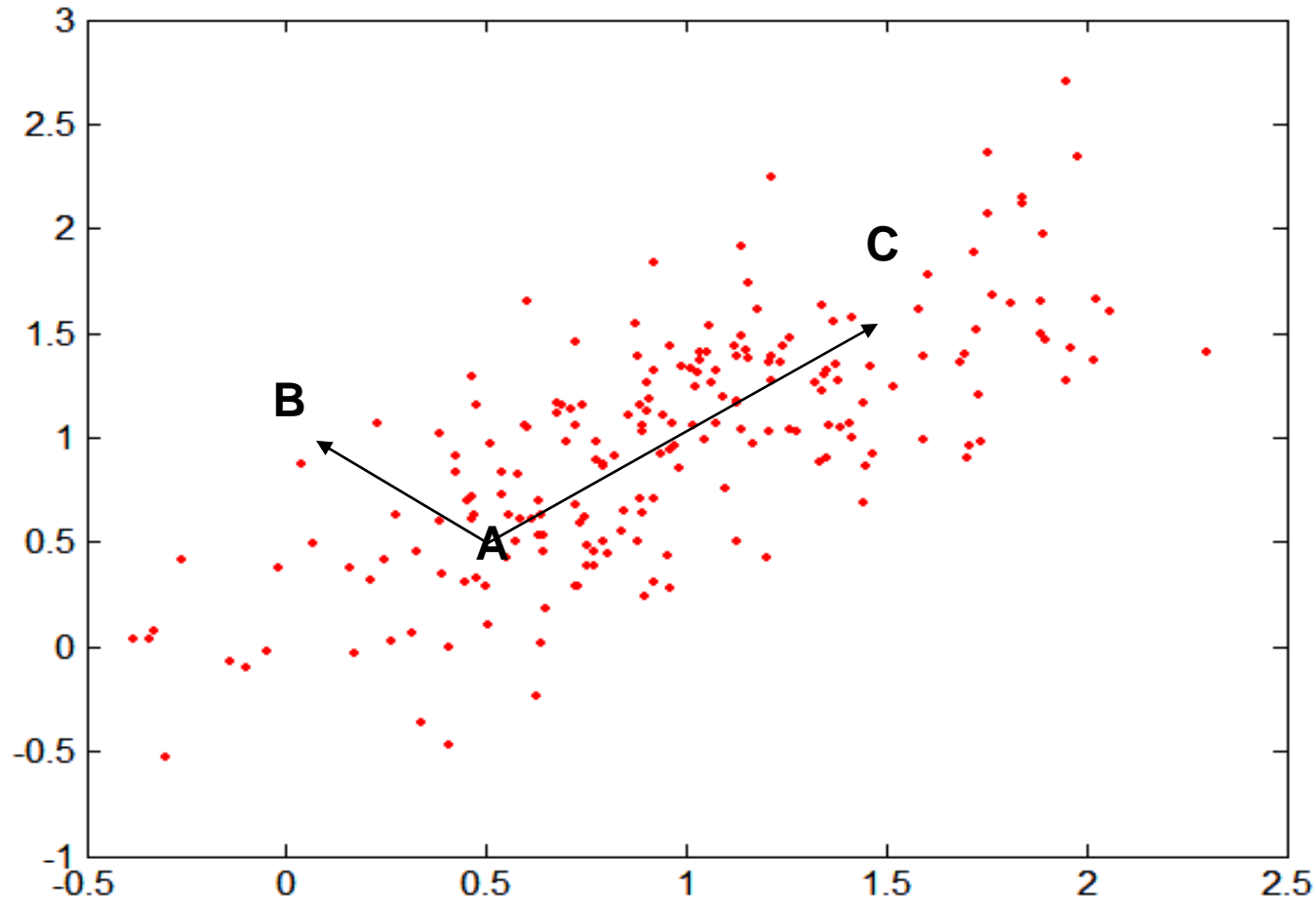
EM algorithm for mixtures

- Alternating algorithm to maximize the likelihood

$$L(\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K; x) = \prod_i \left(\sum_j \pi_j f(x_i; \theta_j) \right)$$

- E-step: compute responsibilities / weights w_{ij} hts, proportional to “probability” (density) point i belongs to cluster j .
- M-step: estimate parameters θ_j .
- Repeat.

Responsibilities / Mahalanobis



Weights related to Mahalanobis distance from each point to center of cluster, balanced with size of cluster.

EM algorithm for Gaussian mixtures

- “Soft” version of K-means, like fuzzy clustering.
- Advantages:
 - Like fuzzy K-means, uses all points to estimate centroids
 - By estimating a covariance matrix per cluster, it can accommodate different sized clusters
 - Being a formal model, information / Bayesian tools can be used to determine K, number of clusters.
- Disadvantages:
 - Like K-means, no guarantee of global convergence
 - If K is large, can be slow.
 - Sensitive to outliers, ill-conditioned covariance matrices.

EM algorithm

- Very powerful tool in statistics.
- Used frequently when there are latent (hidden) variables in the model.
- In mixture models, the class memberships of each point are latent.
- If class memberships known, the problem is easy.
- EM takes advantage of this structure by repeatedly solving the “easy” problem.

Graph-Based Clustering

- Graph-Based clustering uses the proximity graph
 - Start with the proximity matrix
 - Consider each point as a node in a graph
 - Each edge between two nodes has a weight which is the proximity between the two points
 - Initially the proximity graph is fully connected
 - MIN (single-link) and MAX (complete-link) can be viewed as starting with this graph

- In the simplest case, clusters are connected components in the graph.

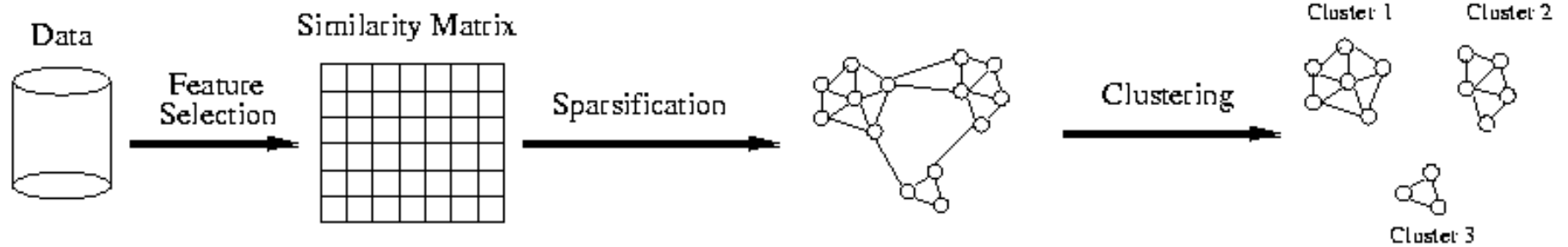
Graph-Based Clustering: Sparsification

- The amount of data that needs to be processed is drastically reduced
 - Sparsification can eliminate more than 99% of the entries in a proximity matrix
 - The amount of time required to cluster the data is drastically reduced
 - The size of the problems that can be handled is increased

Graph-Based Clustering: Sparsification ...

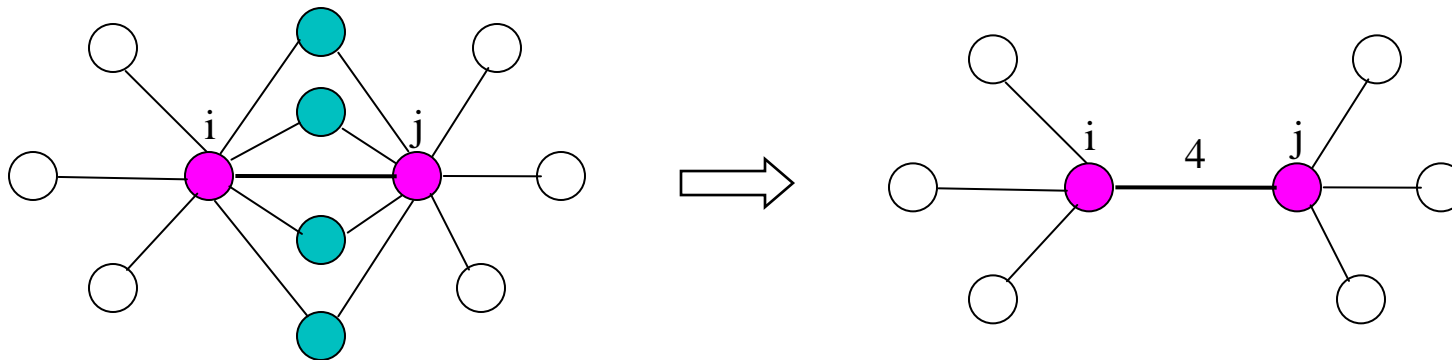
- Clustering may work better
 - Sparsification techniques keep the connections to the most similar (nearest) neighbors of a point while breaking the connections to less similar points.
 - The nearest neighbors of a point tend to belong to the same class as the point itself.
 - This reduces the impact of noise and outliers and sharpens the distinction between clusters.
- Sparsification facilitates the use of graph partitioning algorithms (or algorithms based on graph partitioning algorithms).

Sparsification in the Clustering Process

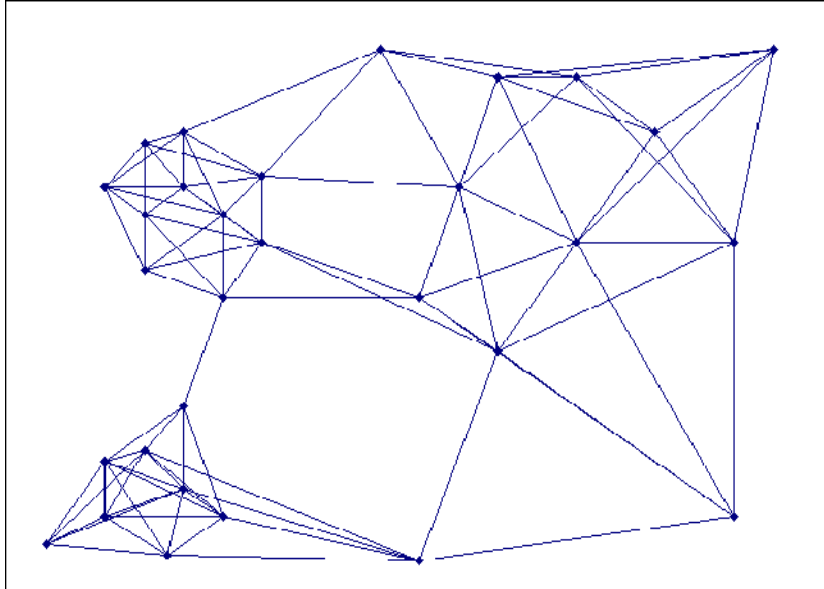


Shared Near Neighbor Approach

SNN graph: the weight of an edge is the number of shared neighbors between vertices given that the vertices are connected

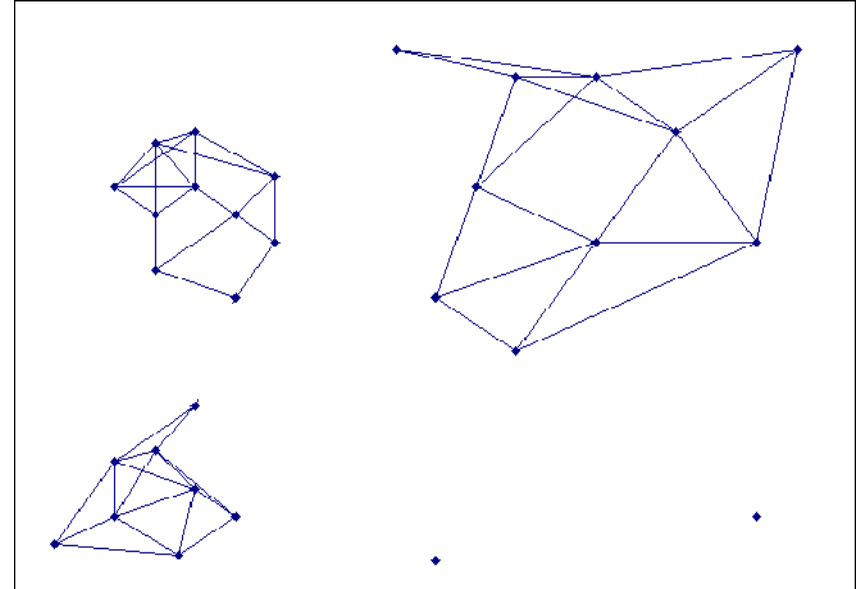


Creating the SNN Graph



Sparse Graph

Link weights are similarities between neighboring points



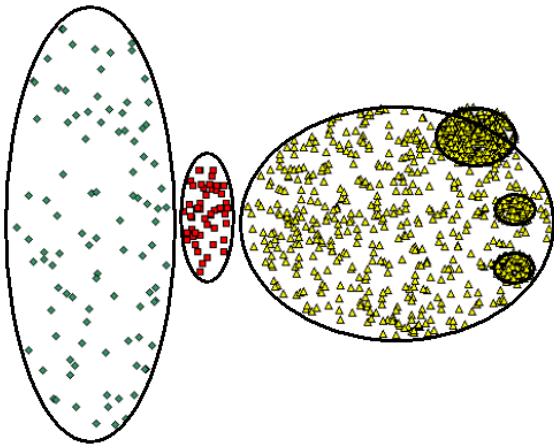
Shared Near Neighbor Graph

Link weights are number of Shared Nearest Neighbors

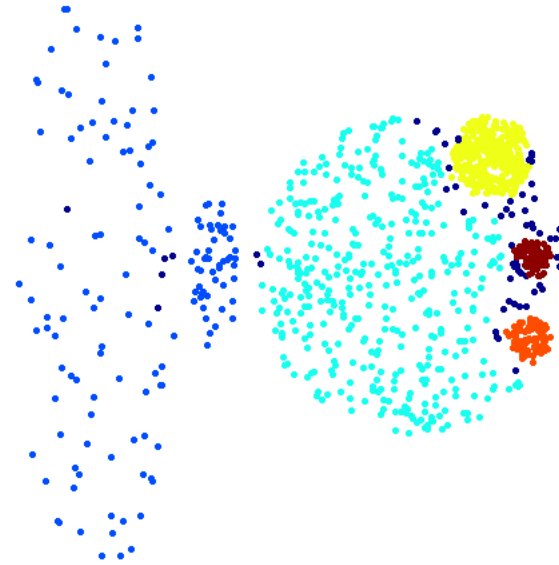
Jarvis-Patrick Clustering

- First, the k -nearest neighbors of all points are found
 - In graph terms this can be regarded as breaking all but the k strongest links from a point to other points in the proximity graph
- A pair of points is put in the same cluster if
 - any two points share more than T neighbors and
 - the two points are in each others k nearest neighbor list
- For instance, we might choose a nearest neighbor list of size 20 and put points in the same cluster if they share more than 10 near neighbors
- Jarvis-Patrick clustering is too brittle

When Jarvis-Patrick Works Reasonably Well



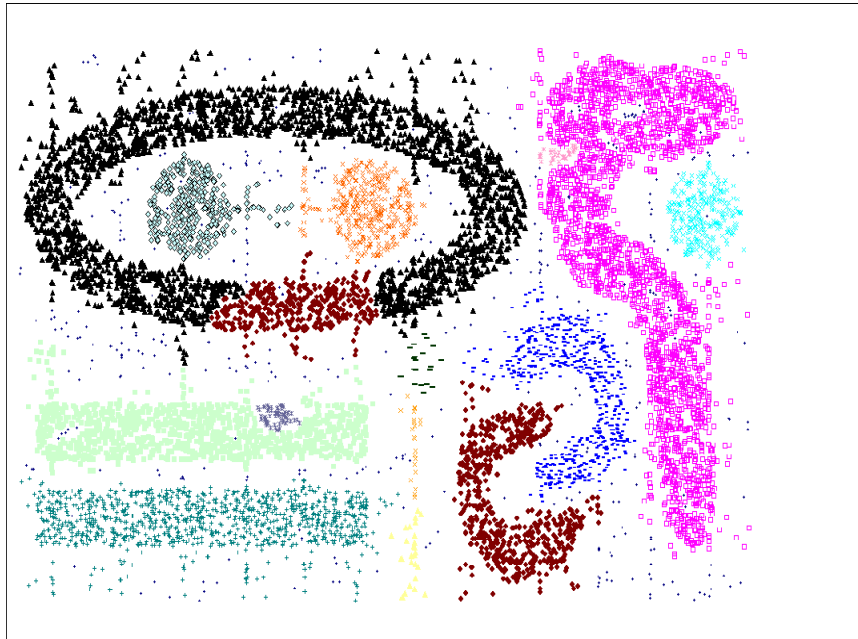
Original Points



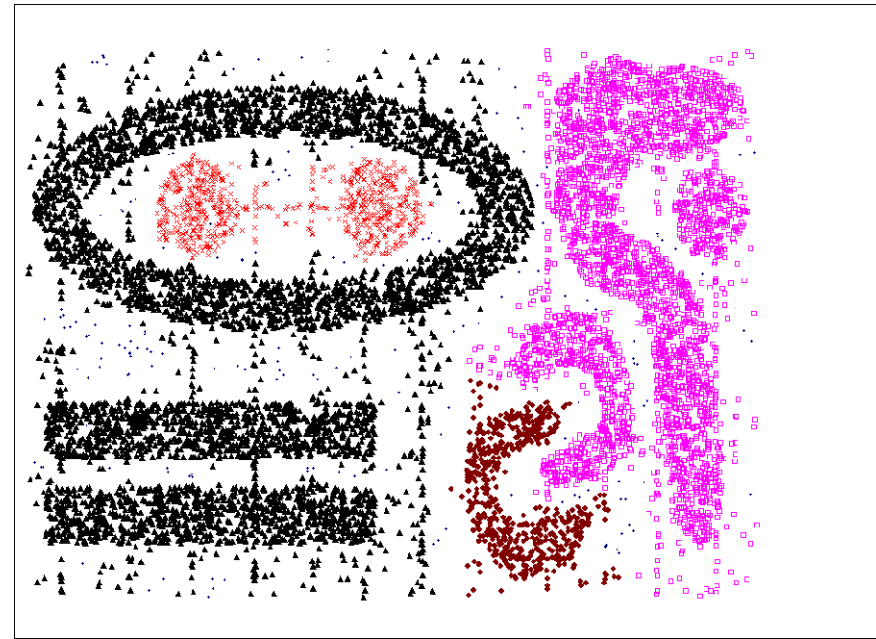
Jarvis Patrick Clustering

6 shared neighbors out of 20

When Jarvis-Patrick Does NOT Work Well



**Smallest threshold, T ,
that does not merge
clusters.**



Threshold of $T - 1$

Spectral Clustering Algorithm

1. **Compute the similarity matrix**

This corresponds to a similarity graph with data points for nodes and edges whose weights are the similarities between data points

$$S_{ij} = \exp(-\text{dist}(x_i, x_j)^2 / 2c)$$

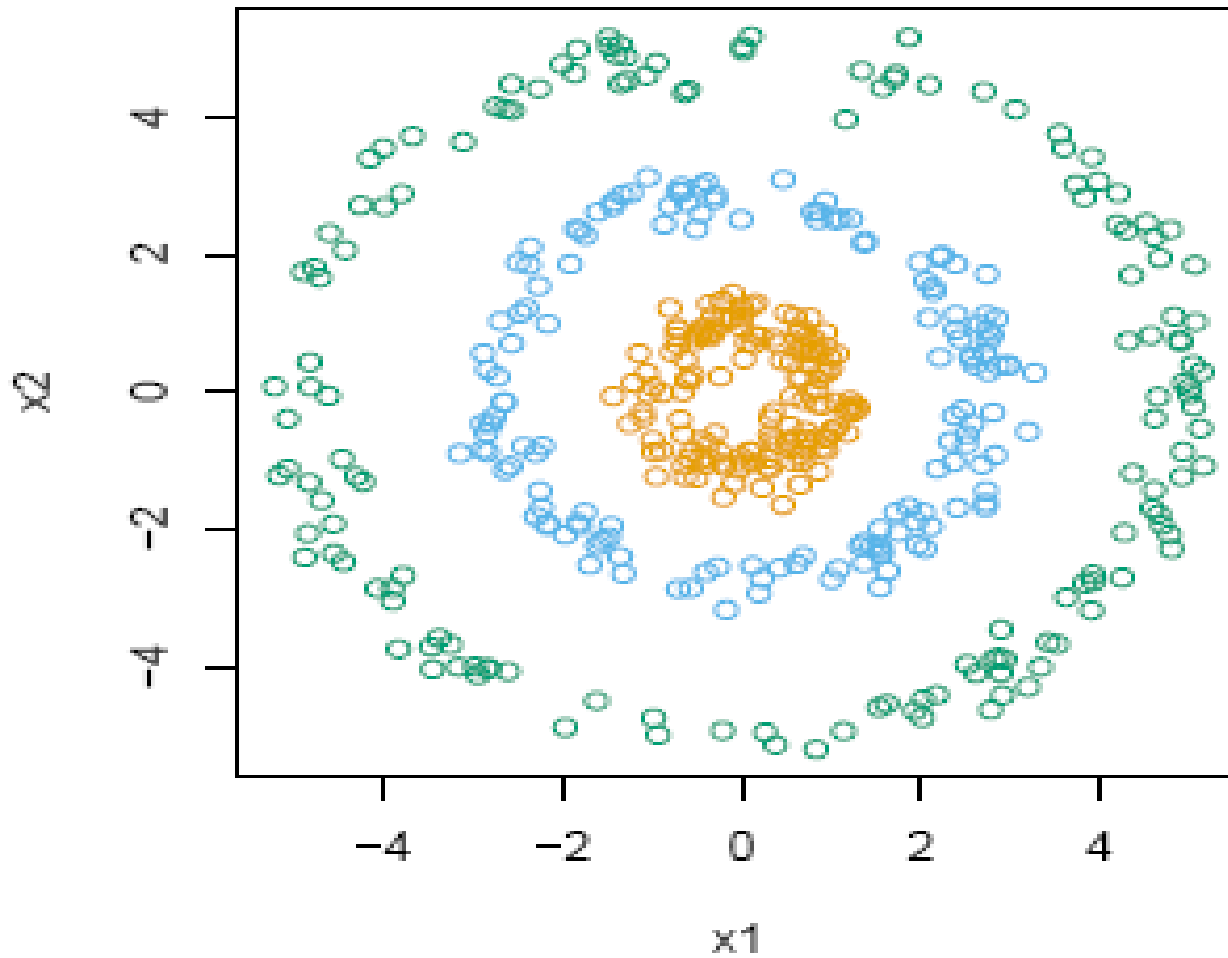
2. **Sparsify the similarity, S , matrix by keeping only the k symmetric nearest neighbors, computing edge weight**

3. **Construct the Laplacian, with D being the row sums of S**

$$L = D - S$$

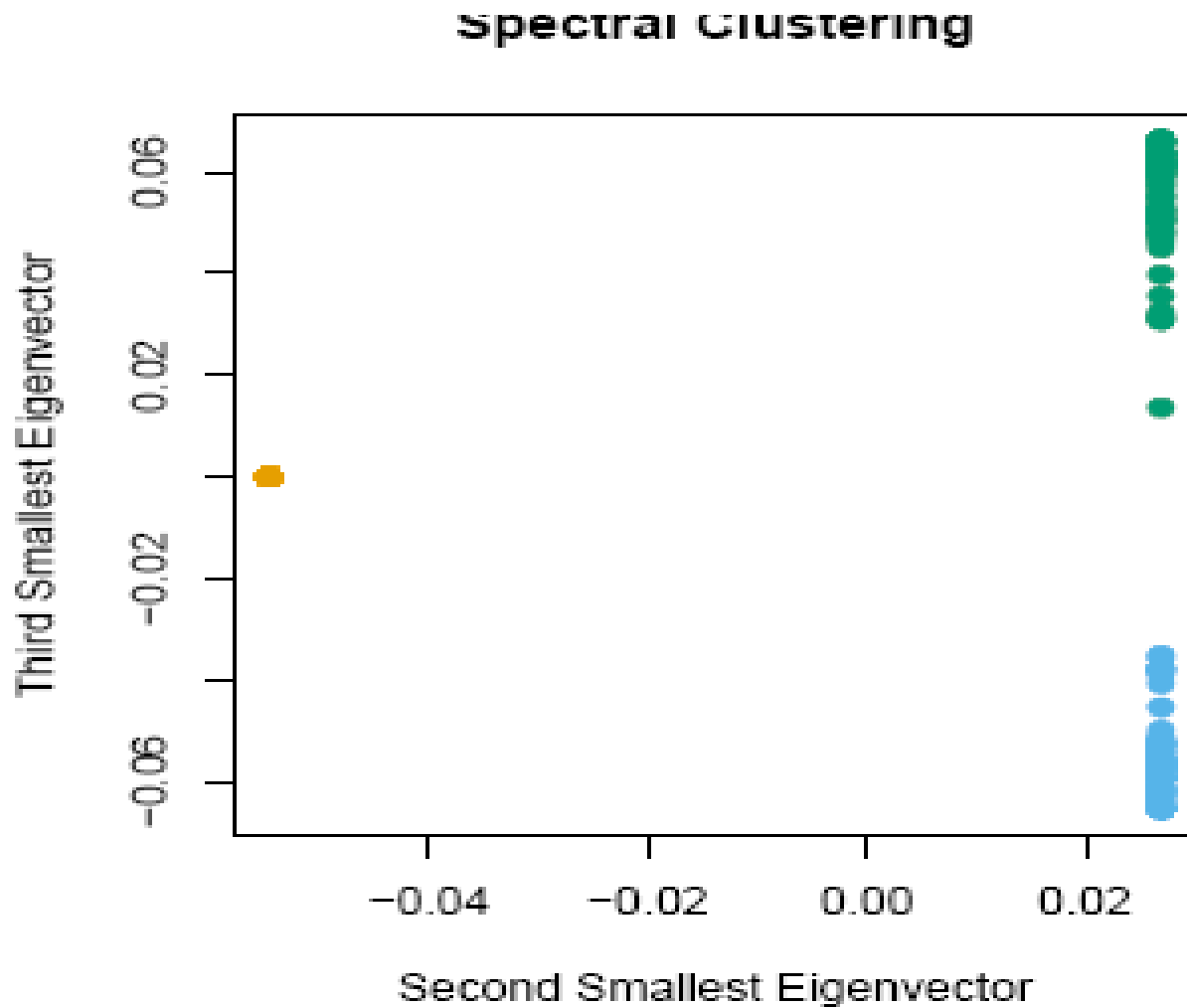
4. **Find the first few eigenvectors of L , and apply K-means to these eigenvectors.**

Spectral Clustering



© Tan, Steinbach, Kumar

Spectral Clustering



Other approaches

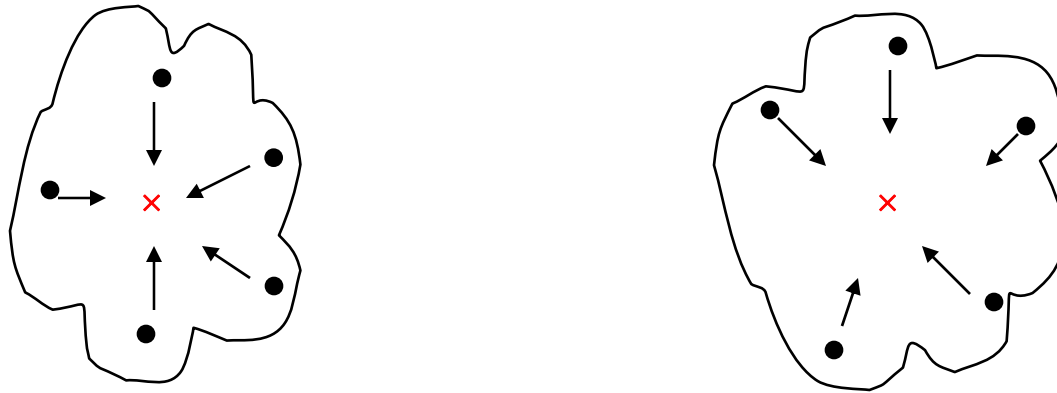
- Spectral clustering creates a new distance between points in dataset.
- Using first few eigenvectors gives a mapping, **M**, from original space to low-dimensional Euclidean space.
- Finally, clustering is applied to Euclidean vectors...
- This mapping, **M** can be created with any similarity / distance, known as **MDS (Multidimensional Scaling)**.
- Book describes related approach: **SOM (Self-Organized Maps)**.

Hierarchical Clustering: Revisited

- Creates nested clusters
- Agglomerative clustering algorithms vary in terms of how the proximity of two clusters are computed
 - ◆ MIN (single link): susceptible to noise/outliers
 - ◆ MAX/GROUP AVERAGE:
may not work well with non-globular clusters
 - CURE algorithm tries to handle both problems
- Often starts with a proximity matrix
 - A type of graph-based algorithm

CURE: Another Hierarchical Approach

- Uses a number of points to represent a cluster

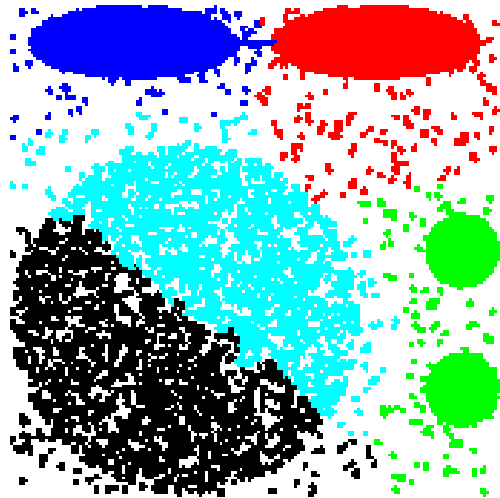


- Representative points are found by selecting a constant number of points from a cluster and then “shrinking” them toward the center of the cluster
- Cluster similarity is the similarity of the closest pair of representative points from different clusters

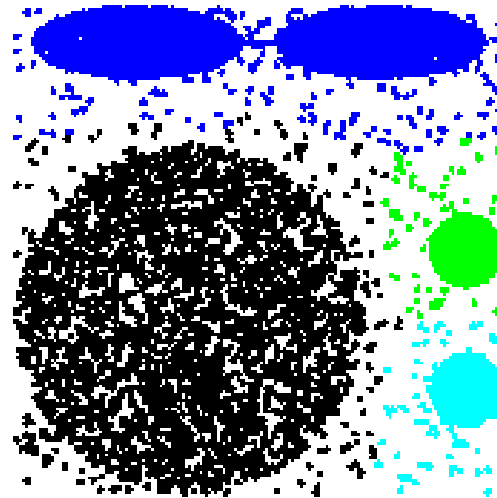
CURE

- Shrinking representative points toward the center helps avoid problems with noise and outliers
- CURE is better able to handle clusters of arbitrary shapes and sizes

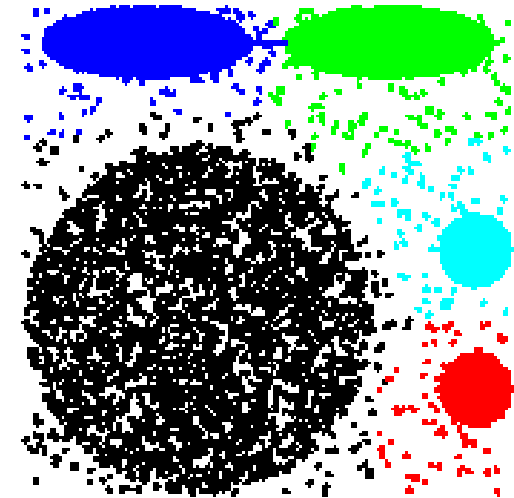
Experimental Results: CURE



a) BIRCH



b) MST METHOD



c) CURE

Picture from *CURE*, Guha, Rastogi, Shim.